DOCUMENT RESUME

ED 447 215 TM 032 123

AUTHOR Morris, Don R.

TITLE Assessing the Implementation of High Stakes Reform: Aggregate Relationships between Retention Rates and Test Results.

PUB DATE 2000-04-00

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Achievement Tests; Educational Change; Elementary Education; *Elementary School Students; *Grade Repetition; *High Stakes Tests; Low Achievement; Student Promotion; *Test Results

IDENTIFIERS *Florida; Reform Efforts

ABSTRACT

The relationship between high stakes testing and retention was studied in Florida in the context of educational reform and controversy over the effectiveness of retaining students in grade. The focus was on the relationship between the percentage of students in grades 1, 3, 5, 7, and 8 falling below the 50th percentile on the Stanford Achievement Test and the percent retained by grade. The scope of the analysis was on the schools years 1986-1987 through 1990-1991, during which the reform went from strong enforcement (high retention policies) to termination. When there was a positive relationship between test results and retention, it tended to occur only in the more affluent schools. Leaving aside the question of whether retention is an effective or desirable way to remediate students and raise educational standards, this study points to the conclusion that high stakes approaches to improving student performance have a low probability of successful implementation. Although the retained-until-remediated policy may be feasible in the better performing schools, cost and space make it expensive even there, and the overall difficulties for implementation are formidable. (Contains 3 tables and 34 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 447 215

Assessing the Implementation of High Stakes Reform: Aggregate Relationships between Retention Rates and Test Results

Don R. Morris
Miami-Dade County Public Schools

Prepared for the
American Educational Research Association
Conference at New Orleans

April, 2000

BEST COPY AVAILABLE

TM032123

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Don Morris
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Assessing the Implementation of High Stakes Reform: Aggregate Relationships between Retention Rates and Test Results

Don R. Morris
Miami-Dade County Public Schools

A new wave of high-stakes reform has begun. The rhetoric has been escalating for some time. At a meeting with the nation's governors on educational standards in March of 1996, President Clinton declared his position to be "No more social promotions" (Cannon, 1996). Such talk marks the start of a new round of acceptability for grade retention policies. From Chicago, the nation's third largest school system, where in 1990 retention was regarded as a "last resort" (Olson, 1990), we hear that "Under the current administration ... the previously condoned policy that academically challenged students should be promoted out of concern for their self-esteem quickly lost favor" (Poe, 1997). Across the country, there is a renewed interest in retention as a major tool in the effort to raise academic standards: "The get-tough stance of holding students back if they cannot show they can do grade-level work has become part of the ongoing movement for tougher academic standards for students nationwide" (Lawton, 1997).

The state of Florida has joined the rest of the nation in a renewed effort to improve the quality of education by means of a policy of high stakes testing. Over the past few years the state legislature has provided for the retention of under-performing students at selected grades and introduced a tough new test, the Florida Comprehensive Achievement Test (FCAT), to measure the progress of both students and schools. The first serious consequences of that new test will be felt at the end of this school year. Its impact has been described recently in the local press:

> [A] new state law . . . cracks down on the practice of social promotion . . . . The new law requires school districts to retain [fourth grade] students who score poorly on the Florida Comprehensive Assessment Test this winter, unless a School Board can show "good cause" for promoting children to fifth grade . . . . State lawmakers banned social promotion in 1997. But the crackdown had no real impact until this year, when they drafted specific rules for promoting children to fifth grade as part of the "A+ Plan" of education . . . . A [Broward County] School Board analysis [has] estimated that . . . 23 percent of all fourth-grade students would have faced retention last year under the new rules. Educators expect larger numbers this year . . . . School Board members overwhelmingly endorsed the notion of holding back low-performing students . . . . In future years, state leaders may require school districts to hold back deficient students in every grade. (de Vise, 1999).

There was an earlier effort to do the same thing. In the 1980s, high retention rates were a national phenomenon of a magnitude that prompted scholars such as Shepard and Smith (1989) to warn of the effects of retention on students. Partly as a result of such reactions, and partly because no tangible positive results from retention were forthcoming, retention as a policy fell out of favor at the end of the 1980s (Olson, 1990).

Florida was at the forefront of that earlier wave of reform also. High stakes testing was strongly applied, and the retention rate was very high throughout most of the 80s. When the reform was

ended by the state in 1990, the elementary and middle grade retention rates dropped to near zero. There are lessons to be learned from this experience that may help in understanding the effects of the current policy trend. The purpose of this paper is to use the past experience of the Miami-Dade County School District – Florida's largest and the fourth largest in the nation – to shed some light on the evaluation and assessment of the reforms being tried today in Florida, and in the nation.

Back-to-Basics Reform in 1980s Florida

The cornerstone of Florida's educational policy from 1978 through 1990 was a strong Back-to-Basics reform, with gatekeeper tests at the third, fifth, eighth, and tenth grades, and a test for graduation that came into use in 1982, having been defended successfully in the courts. The reform relied heavily on retention to raise standards. A state commission in 1990 found that "In elementary and middle schools [in Florida] most remediation takes the form of repeating a grade" (Governor's Commission on the Reform of Education, 1990, p. 49). The retention rates peaked in the Miami-Dade district (until 1998 simply Dade) in the early to mid-1980s (see Morris & Hanson, 1993). The highest rates occurred at the beginning of the educational levels, at 1st, 7th, and 10th grades in the Dade case (Morris, 1991). This reflects a more general pattern, widely occurring and variously reported (see Gottfredson 1988; Karweit 1992; Morris 1993). In addition, the gate tests triggered smaller increases (Morris and Hanson reported small peaks at 3rd and 5th grades).

The objectives of the reform were not met, in either the district or the state. Throughout Florida, the fact seems to be that the high retention policies of the 1980s resulted in little or no demonstrable remediation of at-risk students. The opinion expressed in the report of the Governor's Commission on Educational Reform (1990), and comments by Florida Commissioner of Education Betty Castor (Firing Line, 1992), support this conclusion. For the district, the study of retention in DCPS elementary grades over this period undertaken by Morris and Hanson (1993), and remarks by former DCPS superintendent Joseph Fernandez (Olson, 1990) reflect a lack of success. In 1990, the legislature acknowledged the reform's failure and ended the basic skills testing. In place of the reform, the state began actively to foster increases in promotion and graduation rates (Natale, 1991). The Dade district, which had begun in 1987 to drastically reduce its retention rates at the elementary and middle levels, readily complied. In three years time, from 1987 to 1990, the district moved from a policy that subjected at-risk students to mandatory multiple retentions to one that made every effort to avoid any retention at all.

Where did the reform go wrong? If retention is assumed to be a necessary component of remediation, it is necessary that the retentions bear a strong relationship to student academic performance. While this is always assumed, there is indirect evidence from earlier studies that the assumption did not hold in the Dade district. Since the results of the Stanford Achievement Test then in use showed a clear linear relationship with SES, logic dictated that the retention rate should also bear a linear relationship to SES. However, Morris and Hanson (1993) reported that retentions for the elementary grades (1 through 6) throughout the 1980s were not linearly related to SES (as measured by the percent eligible for Free and Reduced-price Lunch, or FRL). Instead, smoothing techniques revealed that the relationship was curvilinear in virtually every case. The percent retained was positive and linear to about 40 percent FRL, and then leveled out such that there was no relationship between retention and FRL among the poorer schools, who were in the greatest need of remediation.

This curvilinear relationship leads to the conjecture that the reform's program was at best effectively implemented only up to about 40 percent FRL – that is, in the more affluent schools. The loss of a positive relationship beyond that point suggests that high-stakes testing and retention are most closely related where the problems they are intended to resolve pose the least resistance.

Retention and Test Results

The analysis will focus on the relationship between high-stakes testing and retention, controlling physically for SES. This is an aggregate study; the unit of analysis is the school. The concentration is on the relationship between two variables: the percentage of students falling below the 50th percentile on the Stanford Achievement Test (Stanford); and the percent retained by grade (RET). The grades 3, 5, and 8, which were the grades at which a gatekeeper test (known as the SSAT) was given in October (the Stanford was administered in April), and grades 1 and 7, the first years of the elementary and junior high levels, were chosen for scrutiny. The scope of the analysis was the 5 school-years 1986-87 through 1990-91, during which the reform went from strong enforcement to termination.

A Predominantly Nonlinear Pattern

As a preliminary check to ensure that the curvilinear relationship found with respect to FRL transferred to the test results, RET medians across three non-overlapping sub-ranges on the Stanford range were compared by grade by year. The results are shown in Table 1. The table covers the grades 1, 3, 5, 7, and 8. The number of schools varies by grade (schools vary in configuration), and by level (there are many more elementary schools than middle). The numbers in the sub-groupings are also subject to vary from year to year as the percent of students scoring below the 50th percentile varies.

The expected relationship is one in which the larger percent retained occurred in the center of the range. With the exception of seventh grade, the data show the expected pattern. However, all of the years for seventh grade show monotonically increasing medians across the Stanford range, implying linearity. The seventh will be discussed separately below.

For the rest, most of the triplets of medians were of the expected middle-largest pattern. Of the data for the grades 1, 3, 5, and 8 listed in Table 1, fifteen (or 75 percent) of the 20 grade-years have the largest median in the center of the range. Another three (or 15 percent) show a milder curvilinear condition in which the difference between the first and second medians is larger than that between the second and third. Only two (grade 1 in 1989 and grade 8 in 1990) suggest the positive linearity that would imply a successful reform by retention policy, and these occur near the end of the period. These results indicate that the expected curvilinear pattern is prevalent, and justify further analysis.

The next step was to determine whether the pattern was meaningful. Since the smoothing techniques used for the earlier work involving FRL do not lend themselves to significance testing, another approach was used here. The schools were split into two groups, those above and those below 40 percent FRL (where in the earlier studies the deviation from linearity was reported to begin). A comparison was then done of the slopes for the two groups, where the slopes represent the change in the percent retained, for each percent change in students below the 50th percentile of

the Stanford. It was expected that the slopes from linear regression for the groups of schools with less than 40 percent FRL would be positive, and those of the over 40 percent FRL schools would be significantly less positive or even negative. OLS regressions were performed separately on the groups, by grade by year, the slopes compared for the expected pattern, and their differences tested for statistical significance.

Table 1
Median percent retained for schools grouped by non-overlapping segments of the range of the percent of students scoring below the 50th percentile of the Stanford reading comprehension subtest

Grade	Year	Median percents retained Schools grouped by percent of students scoring below 50th percentile					
		10-30%	No. schls	40-60%	No. schls	70-90%	No. schls
1st	1986	5.1	27	9.3	67	8.1	30
	1987	4.4	31	6.3	58	3.2	20
	1988	3.5	22	6.0	75	4.5	21
	1989	1.8	28	3.4	59	5.7	16
	1990	0.0	23	1.1	64	1.1	16
3rd	1986	3.2	14	5.6	49	3.6	35
	1987	3.2	18	3.8	62	2.8	31
	1988	1.3	22	2.4	58	3.2	29
	1989	0.8	24	2.3	60	1.7	26
	1990	0.0	17	0.4	55	0.0	31
5th	1986	1.0	6	2.1	54	1.6	38
	1987	0.5	6	4.2	55	2.6	50
	1988	3.0	13	3.7	63	1.4	29
	1989	0.9	9	2.3	55	1.3	28
	1990	0.0	12	0.8	56	1.0	27
		<45%	No. schls	50-70%	No. schls	75%+	No. schls
7th	1986	4.2	6	8.2	22	11.8	13
	1987	2.6	6	5.8	20	6.6	13
	1988	2.7	6	4.1	24	5.3	12
	1989	1.2	7	2.3	20	3.4	14
	1990	1.5	8	1.5	17	4.5	10
8th	1986	3.7	11	4.9	23	4.3	5
	1987	4.0	9	5.1	19	5.4	10
	1988	3.7	6	3.8	22	3.1	4
	1989	2.4	8	2.5	20	2.0	11
	1990	1.8	9	1.8	28	2.4	6

Let the groups be designated "≤40" for the group of schools with an FRL percentage equal to or less than 40, and ">40" for that group of schools with an FRL membership greater than 40 percent. Table 2 displays the slopes of the two groups for each year, for the grades 1, 3, 5, and 8. The rightmost columns of the table show the differences between the ≤40 and >40 slopes and the p value, representing whether the difference is statistically significant.

Table 2
Comparison of the regression slopes of schools
grouped above and below 40 percent FRL, Grades 1 3, 5 and 8

Grade	Year	Slopes*							
		FRL ≤ 40			FRL > 40			Comparison	
		Slope	No. schls	p	Slope	No. schls	p	diff	p
1st	1986	0.085	65	0.013	-0.004	97	0.914	0.09	0.000
	1987	0.072	63	0.025	-0.041	99	0.228	0.11	0.000
	1988	0.113	55	0.001	-0.011	107	0.722	0.12	0.000
	1989	0.039	51	0.245	0.025	111	0.311	0.01	0.000
	1990	0.009	47	0.201	0.009	115	0.572	0.00	0.727
3rd	1986	0.056	66	0.069	-0.082	94	0.031	0.14	0.000
	1987	0.017	64	0.444	0.004	96	0.884	0.01	0.000
	1988	0.005	56	0.786	0.026	104	0.354	-0.02	0.000
	1989	0.030	52	0.055	-0.014	108	0.487	0.04	0.000
	1990	-0.003	49	0.763	-0.004	111	0.715	0.00	0.028
5th	1986	0.058	66	0.031	-0.033	88	0.198	0.09	0.000
	1987	0.092	64	0.007	-0.077	89	0.073	0.17	0.000
	1988	0.064	56	0.024	-0.087	98	0.002	0.15	0.000
	1989	0.011	52	0.640	-0.033	102	0.230	0.04	0.000
	1990	0.019	49	0.177	-0.004	105	0.845	0.02	0.000
8th	1986	0.059	25	0.160	-0.046	20	0.565	0.11	0.000
	1987	-0.003	25	0.950	-0.111	20	0.190	0.11	0.000
	1988	0.006	21	0.871	-0.022	24	0.692	0.03	0.000
	1989	-0.023	19	0.513	-0.025	26	0.446	0.00	0.534
	1990	-0.034	15	0.385	0.016	30	0.705	-0.05	0.000

*Regression coefficients for Percent Retained regressed on Percent Scoring below 50th Percentile.

Table 2 shows that the slopes reflect the pattern revealed by the medians in Table 1. Three-fourths of the >40 slopes are negative, and 80 percent of the ≤40 slopes are positive. The differences column (diff), obtained by subtracting the >40 slopes from the ≤40 slopes, shows that in approximately 90 percent of the cases, the ≤40 slope is larger than the >40 slope. All but two of the 20 slope differences show p levels well below the traditional 0.05 threshold. All but two of the 18 significant differences are positive. The prevalence of a curvilinear pattern is confirmed.

Retention as a Condition for Remediation

A positive relationship between retention and test scores is assumed to be an aggregate indicator of *potential* remediation, a prerequisite for the success of high stakes reform. Given that assumption, the question becomes one of the extent to which the high-stakes testing has created a condition conducive to successful reform. The final step, then, is to determine which, if any, of the positive relationships identified between retention and the percent of students testing below the 50th percentile are genuine – that is, which differ significantly from zero.

For all grades except seventh, the significance levels for the ≤40 slopes are shown in the fifth column from the right in Table 2. Sixteen of the twenty slopes are positive. Only six, however,

were found to be significantly different from zero at the 0.05 level or better. Those significant positive slopes occur early in the period under analysis (1986-88), and in only two grades, first and fifth.

There appear to be a number of reasons for this result. First, all instances of a significant positive slope were at the elementary level. There are many more schools at the elementary than at the middle level, and the larger Ns contribute to the likelihood that a slope will be significant. Second, the significant slopes occurred in the earlier years of the period under analysis partly because the commitment to the reform was stronger then. In addition, the number of ≤40 schools decreased over time, in part because the build-up of retained students increased the FRL percentages.

Also, the significant slopes were concentrated at two grades. That larger numbers of students were retained in the first grade is due in some degree to a widespread pattern—the retention rate tends to be highest in the first year of a level, and to decrease exponentially through the higher grades of the level (Morris, 1993). The greater overall variance afforded by the higher rate no doubt contributed to the liklihood of a significant outcome. Finally, significant slopes occurred at fifth grade. Most of the emphasis on SSAT results at the elementary level was focused on the fifth grade (as opposed to third, where the test was widely considered to be too easy). There was greater diligence at that grade in ensuring that low scoring students were retained.

For these reasons, it is not clear whether the retention effort in the more affluent schools was generally conducive to successful reform or not. It can be said with confidence only that the six significantly positive slopes indicate that the reform did show some promise with respect to the more affluent schools. It is possible that other of the positive slopes represent meaningful increases, but research must be extended to the student level to make this determination.

For the seventh grade, however, the reform's performance record appears to improve considerably. For that grade, a positive relationship between retention and the percent of students scoring below the 50th percentile was the dominant pattern for both sets of groups, the ≤40 and the >40. Table 3 shows the results for seventh grade.

Table 3
Comparison of the regression slopes of schools
grouped above and below 40 percent FRL, Grade 7

Year	Slopes*										
	FRL ≤40			FRL > 40			Comparison		0 ≤ FRL ≤ 100		
	Slope	No. schls	p	Slope	No. schls	p	diff	p	Slope	No. schls	p
1986	0.134	25	0.073	0.236	20	0.120	-0.102	0.000	0.151	45	0.002
1987	0.056	24	0.195	0.100	21	0.330	-0.044	0.000	0.077	45	0.020
1988	0.042	21	0.378	0.012	24	0.838	0.030	0.000	0.062	45	0.024
1989	0.059	19	0.285	0.044	26	0.516	0.015	0.074	0.042	45	0.129
1990	0.011	15	0.594	0.139	30	0.012	-0.127	0.000	0.060	45	0.006

*Regression coefficients for Percent Retained regressed on Percent Scoring below 50th Percentile.

Table 3 has the same format as Table 2, and in addition displays the slopes for the seventh grade across the full FRL range, along with their p values, in two columns appended on the right. The relationship was uniformly positive across the full range of schools, in every year, and there was only one confirmed instance of downward curvature (1988, where the positive result of the difference in slopes was significant). While the ≤40 schools all showed positive slopes of retention on test results, the >40 schools did also, sometimes having slopes that were more positive. When the two groups were merged, all five slopes were positive across all schools combined, and four of the five (excepting only 1989) differed significantly from zero.

Discussion

While remediation may not follow from retention in any event, it cannot follow from it if students who test poorly are not retained. Leaving aside the controversial question of whether retention is either an effective or a desirable way to remediate students and raise educational standards, this study points to the conclusion that high-stakes approaches to improving student performance have a low probability of successful implementation. Although the retain-until-remediated strategy may be feasible in the better performing schools, cost and space make it expensive even there, and the other reasons reported here imply that the difficulties overall for implementation are formidable. This discussion will consider first how the findings might be accounted for. Then the question of what can be expected of the current reforms is raised. Finally, a possible alternative is offered.

Basic Skills Reform Reconsidered

Four problems of high stakes. Why did it happen that what positive relationship was found between test results and retention tended to occur only in the more affluent schools? The most obvious reasons are cost and classroom space. Simulations based on conditions mimicking those of the Dade district in the 1980s produced estimates indicating that the increase in the overall number of students enrolled in the district due to retention would exceed 8 percent (Morris, 1997), and this would be much higher in low-performing schools. But even where such increases can be managed, as they were to some extent in 1980s Florida with the help of state aid, there are other reasons to be considered. For this study, several are suggested by local experience and/or broader research.

First, considerable pressures were brought to bear on principals. The reason for administering the SSAT early in the school year was to give the information to the teachers, who were then to act upon the results of the gate-tests to provide remediation where needed. Interestingly enough, the main pressure on principals was publication of the SSAT scores in the fall, before there had been any opportunity to act on this information. Principals thus faced strong incentives to manipulate the testing outcomes, rather than provide accurate results.[1] To the extent that such activities occurred, they in turn had the effect of distorting and concealing information from the instructional staff who were to use it in assisting students.

Second, in schools where performance is low (i.e., the high FRL schools), the numbers of students failing to meet the SSAT cutpoint were large, implying that teachers would be expected to retain in large numbers. Over the period 1982-1990, the percent failing to score at or above the cutpoint of the SSAT averaged 15 to 25 percent in those schools where FRL exceeded 70 percent. However, administrators and teachers alike openly acknowledge that teachers who fail too many students risk having their own competence challenged. Presumably, then, there is a reluctance to retain the large

numbers that may be called for, although the research to support that supposition in Miami-Dade does not exist.

Third, teachers often give grades relative to the range of capabilities they find in their classrooms. In high poverty schools, it is not unusual to find a large discrepancy between letter grades and test scores, and district-sponsored research has found this to be the case in the district considered here (Froman, 1992). This systematic grade inflation pattern raises questions and invites challenges that tend to counteract the test results.

Finally, while parents in affluent schools tend to be supportive of their children's school activities in proactive ways, parents from a poverty environment who actively try to support their children tend to do so in reactive ways. Anecdotal evidence gleaned from the experiences of a district-sponsored ombudsman program suggests that these parents tend to oppose negative outcomes such as suspension and retention, in lieu of assisting their children in the achievement of positive outcomes.

The exception and its probable causes. Given these obstacles, why was seventh grade so different? Seventh grades are "special" in a number of ways. It is the grade in which departmentalized instruction begins in earnest, even where the sixth is a middle grade (McPartland, Coldieron, & Braddock, 1987), and it is a grade where peer relationships become dominant and behavioral problems accelerate (Urdan & Maehr, 1995). It is a different and unfamiliar environment also for parents, who are less likely to be acquainted with their children's teachers and the expectations of secondary education. Thus seventh grades are marked by abrupt changes in the roles and relationships of teachers and students, a lack of familiarity among teachers, students and parents, and – partly as a consequence of these things – misbehavior well in excess of elementary school averages.

As a consequence of these differences, the obstacles to retention listed above appear to be altered at seventh grade. Increases in misbehavior add to the disposition of teachers to retain students, skewing the grading curve. The increased disposition to retain reinforces a feeling of mutual support among teachers, helping to override fears of appearing incompetent. At the same time, a lack of familiarity with school, procedures, and teachers, decreases parent/student ability to oppose stronger retention practices. It is possible that only when these factors are all present and interacting that an environment is created that is depersonalized and unstructured enough to make feasible the kind of mass retention that high-poverty schools seem to require if the retention rate is to reflect the test results.

In Florida in the 1980s, these conditions gave middle-level principals considerable leeway for action. There were rumors throughout the reform period to the effect that principals responded to the SSAT at eighth grade (to which particularly strong pressures were applied) by retaining heavily in seventh grade, and promoting "around the test" at mid-year.

Recent Reform Efforts

In January, 2000, on the eve of the first "for the record" administration of the FCAT to the state's fourth graders, the Florida Department of Education anticipated that up to 15 percent would be held back (*Miami Herald*, Jan. 11, 2000). Once again retention is to be the key to remediation. There is a new wrinkle, however. The school as an entity is held responsible if the attending student

membership as a whole fails to meet expected standards. In Florida under the current reform, schools are graded A through F, depending on how their students perform on the FCAT. Schools are "at risk" if a percentage of their students score below a cutoff on the FCAT. Presumably, this threat will force the schools to shape up and fulfill their responsibility.

It is the lowest performing schools that are most threatened. Schools graded F two years out of four find their students eligible for vouchers to attend another school of their choice – public or private. Vouchers have been criticized as a vehicle for "creaming," that is, enticing high performing students to other schools, and leaving the low performers in the failing schools, with substandard instructional faculties (see for example Moe & Shotts, 1995). Although a Florida circuit court recently ruled that publicly funded vouchers to attend private schools violated the state's constitution, the inclusion of the private school option continues pending appeal. With or without private school participation, however, the creaming problem is expected to persist as students of the more concerned parents seek to attend the better performing public schools (see Morris, 1997).

Two other options are available to provide a choice of school for students who wish to leave the school they attend. One is the magnet program/school, which has been widely adopted in Miami-Dade. In this type of school, too, critics have expressed concerns about creaming (Pallas, Natriello, & McDill, 1995). The other, the charter school (a private effort with public assistance), a new alternative in Florida, is also beginning to show the signs of a creaming problem.[2]

Looking at the whole-school-or-nothing choice alternatives, then, there seems to be no reason to think that there will be any more success this time around in dealing with students who are at risk than there was in the 80s. Creaming will leave the low performers where they were. Students will go right on being retained, and the same problems with retention and test scores should apply. Thus there is every reason to think that the problems identified in the 1980s reform will reassert themselves – none have been resolved.

Expanding the Choices

Two major procedures coexist in the Florida school districts at the present time. One; the FCAT testing apparatus, is now in place, and it will be used to sort students (as for example to determine who is retained) no matter what. The second is a widespread policy of separate levels for the basic language arts and mathematics classes based on performance measures beginning in the middle level. Students are promoted into differentiated courses at the next grade that are geared to their level of proficiency. The general term for such practices is tracking, and despite efforts to discourage it, the practice is widespread (Oakes, 1985; Burnett, 1995; Linn, 2000). One source notes that "Tracking . . . remains typical in American secondary schools despite opposition . . . . [and] tracking in various forms has been and remains an important feature of public elementary and secondary education in the United States" (Heubert & Hauser, 1999, pp. 91, 93).

Such a policy is presently in force at Miami-Dade middle schools, where language arts and mathematics courses are divided into three levels by difficulty: Basic, Regular, and Advanced. The decision concerning which level is appropriate is based on the previous year's grades and test scores, and there are good indications that students are in general appropriately placed with respect

to the test scores criterion.[3] There is, however, no choice on the part of students concerning the level at which they are placed.

Thus, while choice of school is presently being broadened, there are few choices available to parents and students with regard to course level. The scope of choice could be increased by shifting the locus of testing, and making entry into the different levels of basic courses contingent upon a voluntary "testing in" at the beginning of the year. Instead of trying to sort students on the basis of last year's grades and test scores (those assessments could still be used to inform students of where they stand), one might make entry contingent on the test scores as an entrance criterion, with decisions concerning testing and preparations left up to the student and family. Summer school and/or tutoring options (public and private) should then be available for all who wish to supplement the previous year's performance with additional preparations (privately produced study aids for the FCAT are already on the market). Only those who decline to be tested would then be placed on the basis of teacher judgment or some similar mechanism.

Addressing the identified problems. These options would address the problems identified in this analysis in the following ways. First, the principal would be relieved of the problem of having to search for creative ways to keep the scores up; the problem of who is tested is taken out of the principal's hands and given to parents. Voluntary testing for placement (individual family responsibility for the test results) would encourage hard work, and the incentives to play games with the numbers tested should be discouraged and minimized by it. The approach would relieve some of the pressure on teachers also, since they no longer must make decisions about promotion. This should simultaneously relieve the pressures for grades, since they are no longer connected to promotion outcomes, while allaying the fear of appearing incompetent in the assignment of disproportionate numbers of low grades, if such are warranted.

Finally, the approach might help less active parents to play a greater role. The earlier observations about the ways in which poor parents try to assist their children suggest that some of the more common efforts to mobilize them may misfire. Educating parents will not alone mobilize them. People are most easily mobilized when faced with a clear and immediate problem for which they perceive a clear solution calling for a clear response.[4] For many poor parents, the expulsion or retention of their child is a clear and immediate problem requiring a clear response - opposition. Linking the test scores to course levels and making the procedures for successful test-taking easily accessible should have a similar effect.

Other concerns. But, assuming that this strategy will adequately address the problems identified by this analysis, what concerns does it raise? Perhaps the most obvious is that it is a form of ability grouping, or tracking. On the one hand, many practicing educators have argued in defense of applying sound but different standards in the form of ability grouping (see Educational Research Service, 1998). Linn (2000) has pointed out that "it is quite possible to have high standards without the standards being common for all students" (p. 10).

On the other hand, even supporters warn, as did Linn, that such support should not "be misinterpreted as supporting placement of some students into tracks with watered-down, basic-skills instruction while others are provided with rich experiences" (p. 11). A major charge against ability grouping is that the low performing get shunted into dead-end basics courses and forgotten (Burnett, 1995; Huebert & Hauser, 1999; Linn, 2000). However, one may argue that the imposition

of high stakes can itself be interpreted as a form of de facto tracking. Retention, as the primary methodology of such reforms, can be seen as a kind of "ability grouping device" in which low-performing students are forced into an overage-in-grade group sharing failure as a common experience around which to form a peer identity. A large body of research on the effects of retention on student achievement has accumulated over the past 20 or 30 years, and the overwhelming conclusion of that work is that there has been no positive effect. As one report put it: "no other topic of educational research has generated a conclusion so unanimous as that generated by this burgeoning body of evidence [on retention]" (Texas Education Agency, 1993, p. 29).

The concern that some students might be denied equal quality of education is in fact a restatement of the creaming problem, and this risk is shared with the popular alternatives now in favor – vouchers, charters, and magnets. Putting testing for placement in the hands of parents addresses that concern by leaving mobility between tracks open, with students informed about what options are available to them for seeking to improve their position. With voluntary testing for placement, ability grouping may be less objectionable when compared to probable consequences of the alternatives: branding schools as failures, closing schools, and whole-school-or-nothing choices.

A Broader Perspective

In Florida and nationally, there have now been two high stakes reforms with elevated retention rates in the past 20 years or so, interspersed with an anti-retention counter-reform. A number of researchers have noted that retention and social promotion tend to alternate (e.g., Karweit, 1992), but this cycle of retention/social promotion as a long-term pattern of educational reform has been largely ignored by analysts. It seems likely that the cyclical pattern is related to the pressures being constantly brought to bear on educational systems. Linn (2000) has described what may be one of the mechanisms of the reform cycle. Noting that when a new test is introduced, low initial scores are followed by increases for a few years before a leveling off sets in, Linn makes the following observation:

> Policymakers can reasonably expect increases in [test] scores in the first few years of a program . . . with or without real improvement . . . . The resulting overly rosy picture that is painted by short-term gains . . . gives the impression of improvement right on schedule for the next election. (p. 4)

Although Linn did not pursue the scenario to its logical conclusion, it follows that if the gains do not reflect genuine achievement the pressures invariably reappear for more change later on, furnishing a salient anti-testing issue for later elections. Thus, in a sense, high stakes is built into the cycle.

The four problems with standards reform identified in this analysis result from high stakes pressures. It is due to those pressures that principals are tempted to conceal and distort information from the testing, that teachers are encouraged to counteract and/or ignore test information, and that parents and students are often motivated to oppose the consequences. Taken singly or together, these actions interfere with the normal feedback of accurate information about student performance to educational decision makers, and it has been argued that this impaired feedback loop is the fundamental cause of the cycling between retention and social promotion (Morris, 1994, 1996). To the extent that relieving the pressures of high stakes reform will result in a more open reporting of actual outcomes, reform policies should become more amenable to successful planning and management. That is what the recommendations put forth here are intended to achieve.

Notes

Author's note: The work reported in this paper is not related to my duties for the Miami-Dade County Public Schools, and the district bears no responsibility for the contents.

[1]The pressures resulting from high-stakes testing are often accompanied by reports of unauthorized activities, in the Dade district as elsewhere. Most remain only rumors. Occasionally, however, there is documentation confirming that such activities have taken place. See for example Allington and McGill-Franzen (1992), who have reported evidence of such activities in New York state. More generally, Raspberry (1999) writes that the problem is serious enough that a House subcommittee is looking into allegations concerning recent National Assessment of Educational Progress tests.

[2]Charters are recent enough that research on their performance is scarce, but there is already some evidence of creaming. As an example, one elementary charter in a minority neighborhood in Miami-Dade, which has shown early achievement results somewhat better than the neighborhood average, boasts the same ethnic makeup as comparable regular schools, but an FRL statistic that is 12 to 23 percentage points lower (Hanson, 2000).

[3]A reanalysis of Froman's (1993) data reveals that in both low and high performing schools, the mean Stanford scale score of students at each more difficult level of a course is progressively higher.

[4]In an early work (1957), James S. Coleman noted that some kinds of problems, such as floods, have the effect of uniting a community. Under such conditions the actions to be taken are clearcut (filling sandbags, etc.) and the perceived solutions clear and "do-able." Other problems, such as droughts, have no clear solutions or actions that logically follow, and lead to conflict, disharmony, and uncoordinated activities. The predicament of poor parents trying to assist their children is directly analogous. Proactive responses – day-to-day encouragement and help with homework, coupled with regular school-home interaction – require planning and foresight and are not often clearly linked to outcomes. In particular, proactive behavior is difficult to link directly to outcomes removed in time, such as promotion and retention, and is hard to motivate. An expulsion or retention provides a clear target, crystallizing a goal to be overcome by well understood methods of protest.

References

Allington, R. L., & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York. *Educational Policy, 6*, 397-414.

Burnett, G. (1995). *Alternatives to ability grouping: Still unanswered questions*. Digest No. 111. New York: ERIC Clearinghouse on Urban Education.

Cannon, A. (1996, March 28). Governors talk tough on school standards. *The Miami Herald.*

Coleman, J. S. (1957). *Community conflict*. New York: The Free Press.

de Vise, D. (1999, November 24). Broward schools brace for crisis. *The Miami Herald.*

Educational Research Service. (1998). *Ability grouping: Effects and alternatives*. Arlington, VA: author.

Firing Line. (1992, January 16). *A report card for schools*. Telecast transcript #927, distributed by Southern Educational Communications Association, Columbia, SC.

Froman, T. (1993*). On the relationship between standardized test scores and classroom grades*. Miami, FL: Dade County Public Schools, Office of Educational Accountability.

Governor's Commission on the Reform of Education. (1990). *Reforming education in Florida*. Tallahassee, FL: Author.

Gottfredson, G. D. (August, 1988). *You get what you measure - You get what you don't: Higher standards, higher test scores, more retention in grade*. Center for Research on Elementary and Middle Schools. Report 29. Baltimore: The Johns Hopkins University.

Hanson, M. K. (2000). *MDCPS 1998-99 charter schools review*. Office of Evaluation and Research, Report No. PLCY-00-01. Miami, FL: Miami-Dade County Public Schools.

Heubert, J. P. & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Karweit, N. L. (1992). Retention policy. In M. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., pp. 1114-1117). New York: Macmillan.

Lawton, M. (1997, June 11). Promote or retain? Pendulum for students swings back again. *Education Week*, pp. 1, 37.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*, No. 2, 4-16.

McPartland, J. H., Coldiron, R., & Braddock, J. H. (1987). *A description of school structures and classroom practices in elementary, middle, and secondary schools*. Center for Research on Elementary and Middle Schools. (Report 14). Baltimore: The Johns Hopkins University.

Miami Herald. (2000, January 11). Up to 15 percent of state's fourth-graders might flunk. *The Miami Herald*.

Moe, T. M. & Shotts, K. W. (1995). Computer models of educational institutions: The case of vouchers and social equity. In R. L. Crowson, W. L. Boyd, and H. B. Mawhinney (Eds.), *The politics of education and the new institutionalism*. Washington, DC: Falmer.

Morris, D. R. (1996). Institutionalization and the reform process. *Educational Policy, 10*, 427-447.

Morris, D. R. (1994, November). *Is real progress in education being concealed by the operation of the reform process? Some insights from a system dynamics simulation*. Presented at the Florida Educational Research Association conference in Tampa.

Morris, D. R. (1993). Patterns of aggregate grade-retention rates. *American Educational Research Journal, 30*, 497-514.

Morris, D. R. (1997). *A simulation model for intradistrict public school choice*. Office of Evaluation and Research, Report No. PLCY-97-02a. Miami, FL: Dade County Public Schools.

Morris, D. R. (1991, April). *Structural patterns and change in grade retention rates: An aggregate analysis of data from a large urban school district, 1982-1989*. Presented at the American Educational Research Association Conference, Chicago, Illinois.

Morris, D. R. (1997). *Structural Consequences of Retention Policies: The Use of Computer Models to Inform Policy*. Presented at the Florida Educational Research Association Conference, Orlando.

Morris, D. R., & Hanson, M. K. (1993). Consequences of reform: Retention rate fluctuations in Dade County. In S. Bacharach & R. Ogawa (Eds.), *Advances in research and theory of school management and educational policy* (Vol. 2, pp. 121-157). Greenwich, CT: JAI Press.

Natale, J. A. (1991, January). Promotion or retention? *The Executive Educator*, pp. 15-18.

Pallas, A. M., Natriello, G., & McDill, E. L. (1995). Changing students/changing needs. In E. Flaxman & A. H. Passow (Eds.), *Changing populations, changing schools*: Ninety-fourth yearbook of the National Society for the Study of Population. Part II. Chicago: National Society for the Study of Education.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Olson, L. (1990, May 16). Education officials reconsider policies on grade retention. *Education Week*. [On Line.] Available: http://www.edweek.org/ew/vol-09/09400044.h09

Poe, J. (1997, July 1). For students, a summer of drills. *The Chicago Tribune*.

Rapsberry, W. (1999, June 7). The low road to better test scores. *The Washington Post National Weekly Edition*.

Shepard, L., & Smith, M. (Eds.). (1989*). Flunking grades: Research and policies on retention*. New York: Falmer Press.

Texas Education Agency. (1993). *Closing the gap: Acceleration vs. remediation and the impact of retention in grade on student achievement*. Austin, TX: author.

Urdan, T. C. & Maehr, M. L. (1995). Beyond a two-goal theory of motivation and achievement: A case for social goals. *Review of Educational Research, 65*, 213-243.